EMBO
reports

concept

Towards completion of the Earth's proteome

Carolina Perez-Iratxeta[1], *Gareth Palidwor*[1] *& Miguel A. Andrade-Navarro*[1,2,3+]

[1]Ottawa Health Research Institute, Ottawa, and [2]Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada, and [3]Max Delbrück Center for Molecular Medicine, Berlin, Germany

New protein sequences are deposited in databases at an accelerating pace; however, many of these are homologous to known proteins and could be considered redundant. If all historical releases of the protein database are analysed using the original sequence-clustering procedure described here, the fraction of newly sequenced proteins that are redundant is increasing. We interpret this as an indication that the sequencing of the Earth's proteome—the complete set of proteins on Earth—is approaching completion. We estimate the approximate size of the Earth's proteome to be 5 million sequences, most of which will be identified during the next 5 years. As the Earth's proteome nears completion, cluster analysis of the protein database will become essential to identify under-explored taxa to which future sequencing efforts should be directed and to focus research on protein families without experimental characterization.

Keywords: protein sequence database; genomics; sequencing project; database annotation; phylogenetic analysis

EMBO *reports* (2007) **8,** 1135–1141. doi:10.1038/sj.embor.7401117

Introduction

The collection of new protein sequence information from any organism increases our understanding of the biology of that organism and of all organisms bearing evolutionarily related protein sequences. It is therefore useful to sequence and annotate as many different protein sequences as possible. This is beneficial to all methodologies and fields of research that take advantage of large numbers of protein sequences to provide a better understanding of biology, such as evolutionary biology (Pace, 1997), phylogenetic profiling (Huynen & Bork, 1998) and industrial biocatalysis (Schmid *et al*, 2001). The completion of the Earth's proteome—determining the set of distinct protein sequences from all of Earth's organisms—is a desirable objective that involves a significant part of the scientific community.

Today, most protein sequences are inferred from gene sequences obtained with DNA-sequencing technologies. These technologies are rapidly evolving towards faster and cheaper methods to, for example, sequence individual genes (Sanger, 2001), obtain expressed sequence tag (EST) libraries from organisms (Adams *et al*, 1992), sequence entire genomes—providing the complete repertoire of gene functions of an organism (Koonin & Mushegian, 1996)—or sequence DNA from environmental samples (Rondon *et al*, 2000; Tyson *et al*, 2004; Venter *et al*, 2004; Yooseph *et al*, 2007). Consequently, the rate of deposition of protein sequences has accelerated (Fig 1) and the largest databases such as Entrez (Wheeler *et al*, 2007) and UniProt (Bairoch *et al*, 2007) now contain millions of sequences.

To sequence the Earth's proteome completely could seem an impossible task, as the number of species is estimated to be in the order of millions (Torsvik *et al*, 2002). It is likely that the rate of production of new gene sequences in nature exceeds any practically achievable sequencing rate. For example, it has been estimated that every gene broadly shared by the prokaryotic population would suffer a mutation somewhere on Earth in a timescale of minutes (Whitman *et al*, 1998). However, it is obvious that not all of the differences between similar protein sequences have equal biological significance. Indeed, a cursory observation of the protein sequence databases shows that they contain many sequences that are similar to each other. Many of those are versions of the same protein obtained from phylogenetically related species and probably perform equivalent functions (Fitch, 1970).

In this article, we explore the concept that by sorting all known protein sequences into groups of sequences similar over their full length, it is possible to estimate the number of distinct protein functions within already sequenced proteins. Then, assuming that the number of total protein functions on Earth does not vary appreciably on a scale of years, we hypothesize that completion of the Earth's proteome would be reached in practical terms when any newly sequenced protein is already represented in the database by another functionally equivalent protein.

In 2001, Vitkup and colleagues (Vitkup *et al*, 2001) proposed that, as more genomes are sequenced, additional sequencing will eventually result in proteins that are increasingly redundant to known proteins. In this Concept, we explore the idea that such an increasing trend in the redundancy of the sequences added to the database might be used to infer the size of the Earth's proteome and the time needed to fully sequence it. Two recent studies concluded that such an increase in redundancy is not yet happening. Marsden and coworkers analysed the increase in protein clusters using 169 complete genomes available in 2005 (Marsden *et al*, 2006). They plotted

[1]Department of Molecular Medicine, Ottawa Health Research Institute, 501 Smyth Road, Ottawa, Ontario K1H 8L6, Canada
[2]Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H 8M5, Canada
[3]Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany
[+]Corresponding author. Tel: +1 613 737 8899 Ext. 73135; Fax: +1 613 739 6294;
E-mail: mandrade@ohri.ca

Submitted 9 July 2007; accepted 15 October 2007

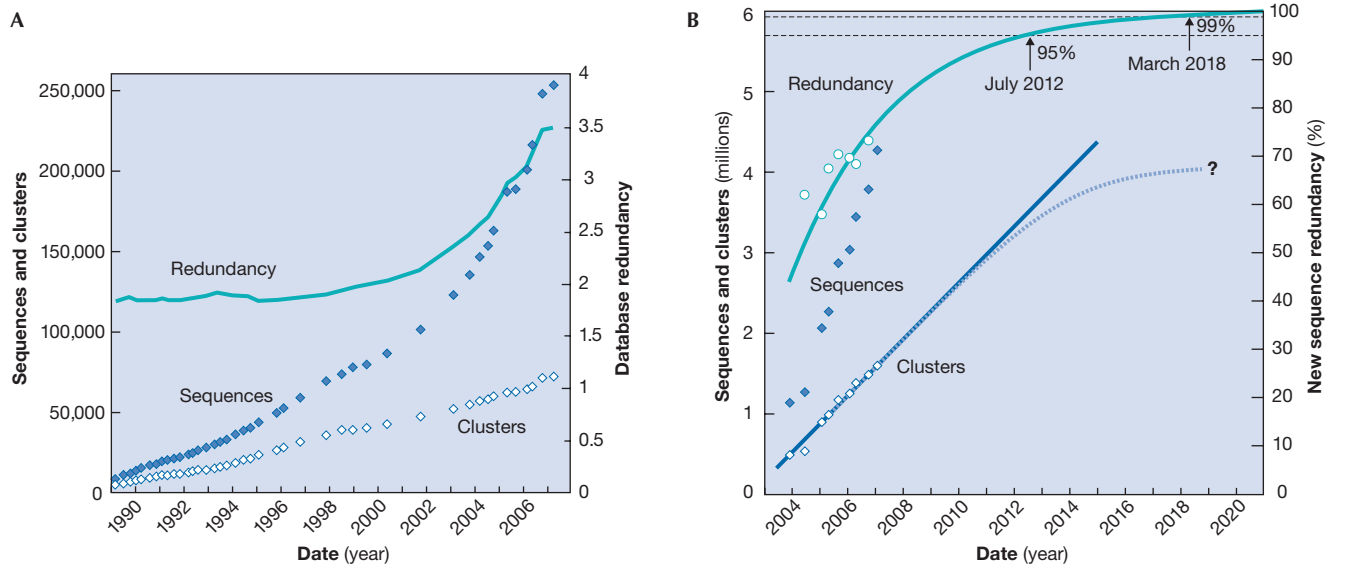

Fig 1 | Analysis of sequencing trends. **(A)** Historical evolution of the SwissProt database. Filled diamonds represent the number of sequences and open diamonds represent the number of sequence clusters. The continuous line is the database redundancy, which is calculated as sequences divided by clusters. Although sequences are added at increasing speed, the number of clusters increases linearly. As a result, the database redundancy increases. **(B)** Extrapolation of sequencing trends in UniRef100. Filled diamonds represent the number of sequences in UniRef100, open diamonds represent the number of sequence clusters (the cluster data can be adjusted to a line) and open circles represent the percentage of sequences new to a version of UniRef100 that clustered with sequences present in the previous version of the database. The redundancy data can be adjusted to an asymptotic function of the form $g(x) = 56 \times (1 - \exp(bx)) + 44$ for $b = -0.0235735$, where x is the number of months since release of UniRef100 version 1 (December 2003). Redundancy of new sequences at 95% is expected for the year 2012, and at 99% for 2018. A high estimate of 5 million sequences is proposed as the size of the Earth's proteome, assuming that the discovery of new protein clusters will start to slow (discontinuous line with a question mark).

the number of protein clusters in the cumulative set of proteins formed by the first *n* genomes sequenced. Their conclusion was that the number of known protein clusters was at the time continuously expanding. However, they did not take into account the variable timescale on which these genomes were sequenced. More recently, Raes *et al* simulated the growth in number of protein clusters when proteins from a pool of more than 2 million sequences—all proteins from complete genomes and proteins deduced from four large sets of environmental samples—are successively taken at random (Raes *et al*, 2007). Again, their conclusion was that clusters grow linearly as more proteins are sequenced. Neither of these studies considered the real-time evolution of the complete protein sequence database. There are good reasons for this, as both groups used variations of the Markov cluster algorithm that required an 'all against all' sequence comparison computationally unfeasible to run on all versions of the protein database.

In our opinion, a realistic analysis of the sequencing trends requires the study of the complete protein data set, including not only the current version, but all its historical versions. In this Concept, we explore a pragmatic clustering that compares only proteins of similar length—under the assumption that proteins of very different length will have different functions—thus reducing enormously the amount of computation required. This clustering method has been specifically produced for the purposes of this analysis, which is to study the evolution of the database and not to produce inclusive clusters. This method shows the increase in redundancy of newly produced sequences and allows us to make the first estimate of the total of protein functions on Earth. In addition, this clustered representation of the protein database can be used to direct the sequencing of new genomes and to focus experimentation on large uncharacterized protein families.

Status of the completion of the Earth's proteome

An original sequence clustering algorithm was used (see supplementary Methods section online) to study the evolution of the protein database along time. If SwissProt—the oldest actively maintained protein database (Bairoch & Boeckmann, 1991)—is analysed, one can see that whereas the number of its sequences has increased at an accelerating pace since 1995, when the first genome projects started (Fleischmann *et al*, 1995; Fraser *et al*, 1995), redundancy also increased and is now at a level of 3.5 sequences per cluster in SwissProt v52 (October 2006; Fig 1A). A similar trend can be observed in the less curated but larger UniRef100 (UniProt Reference Clusters) database (Suzek *et al*, 2007). Fig 1B represents the number of sequences and clusters, as well as the percentage of sequences added between consecutive versions of this database that cluster with some sequence of the previous database version—and therefore could be considered to represent proteins redundant with known proteins. For example, 73% of all sequences added to the UniRef100 database between releases 9.0 (October 2006) and 10.0 (February 2007) clustered with previously deposited sequences. Following the hypotheses presented above, it is reasonable to assume that, as completion of the Earth's proteome approaches, the fraction of new sequences

redundant to known proteins will asymptotically approach 100%. Assuming that the current trends of new sequence deposition continue, it is possible to adjust the historical percentages of redundant proteins to a time-dependent asymptotic function and estimate that, by 2012, 95% of the sequences added to the database would be redundant to those already in the database (Fig 1B). The number of clusters (distinct proteins) is now growing linearly, suggesting a total of 6 million clusters for 2017. However, it is reasonable to assume that this growth will decrease and a value of 5 million for the total number of distinct proteins on Earth is proposed here as a high estimate. On the basis of this, approximately 32% of all distinct proteins on Earth are already known.

Examination of sequencing coverage by taxa

These results suggest that the sequences of the majority of proteins on Earth will be obtained in the next 5 years. But is this a reasonable prediction? On the one hand, the production of sequences will probably keep growing as it has been doing in recent history (Fig 1); on the other hand, obtaining biological sources of the remaining unknown sequences will be increasingly difficult. This is becoming evident for prokaryotes, as there are many species that cannot currently be cultured or might live in environments difficult to access or even to imagine. The technical difficulties of obtaining novel samples can be solved with imaginative approaches—such as PCR amplification of genomes from single cells (Ottesen *et al*, 2006). However, current protein sequence data must be examined to determine branches of the taxonomic tree likely to contain organisms that have proteins with functions not yet represented in the proteome. This Concept proposes a clustering method and a protein cluster database that can be used to identify such under-explored taxa. For each cluster, the taxon common to all lineages of the sequences in that cluster is computed. Then, the sizes of the clusters from particular taxa can be studied. Well-explored taxa tend to have clusters with many proteins, as sequencing from related organisms produces similar versions of the same protein—for example, several *Mycobacterium* species have now been sequenced—whereas under-explored taxa have clusters with average sizes closer to one.

An overview of the current taxa distribution for the 1.35 million clusters obtained from UniRef100 (release 8.5, September 2006) is represented in Fig 2, in which the boxes represent taxa, the number of associated clusters is represented by box size, and the average number of sequences per cluster is represented by colour intensity. The complete data are provided in Table S1 available at: http://www.ogic.ca/projects/clusters/sorted_allcluster_taxonomy_8.5.zip. By kingdom, the number of eukaryotic clusters (688,850) is slightly larger than that of bacterial clusters (543,178), with a minor contribution from archaeal ones (31,329). Visual inspection of Fig 2 shows a smaller number of larger partitions for Eukaryota than for bacteria, reflecting the fact that fewer eukaryotic organisms have been sequenced but that their genomes are evidently larger. The average size of the eukaryote-specific clusters—1.38 million sequences for 0.69 million clusters, two sequences per cluster—is slightly lower than that for bacteria-specific genes—1.2 million sequences for 0.54 million clusters, 2.4 sequences per cluster—indicating the higher level of sequencing coverage reached for prokaryotic organisms. The low level of sequencing for archaeal organisms is reflected in both the small number of sequences in the database, and in the small average size for archaea-specific clusters—1.54 sequences per cluster. This confirms that the archaeal kingdom is largely unexplored. Within each kingdom, the large variability in average cluster size across the different taxa indicates that some taxa are much better explored—for example, Mammalia has 1.71 sequences per cluster, and Bacilli has 1.91—than others—such as Alveolata, which has 1.29, Echinodermata, 1.08 or Deltaproteobacteria, 1.17. These data indicate directions for future sequencing efforts that might yield high proportions of new protein functions.

The annotation of the Earth's proteome

The sequencing of the Earth's proteome can be viewed as a global genome-sequencing project. An important and challenging part of any genome-sequencing project is the characterization of the encoded proteins (Casari *et al*, 1995) and, in this respect, the Earth's proteome is the same. Many large-scale efforts are ongoing to assess protein structure ('structural proteomics'; Gaasterland, 1998), functional domains (Finn *et al*, 2006; Letunic *et al*, 2006) and function ('functional proteomics'; Adam *et al*, 2002); and many genes have already been characterized experimentally to some degree as described in the biomedical literature (MEDLINE; Wheeler *et al*, 2007). The clustered database described above can be used to assess and focus the progress of the annotation of the Earth's proteome in all of these categories. Each protein in the UniRef database might have associated information on structure (by links to the Protein Data Bank database), bibliography (by links to MEDLINE), function (through controlled vocabulary annotations like Gene Ontology terms) and domains (by links to the Pfam domain database). For each of these properties, the fraction of annotated proteins in the database gives a sense of the progress in completing the annotation for that type of information (column 1 in Table 1). The assumption that sequences clustered by full-length similarity are mutually characterized by any of the members of the cluster is at the basis of most of the use of bioinformatics by the biomedical community (Perez-Iratxeta *et al*, 2007). Therefore, the number of clusters with annotated members gives insight into the usefulness of annotations in characterizing the proteome (column 2 in Table 1). For example, it is generally more informative to know the structure of two proteins in different clusters than to know two structures corresponding to two proteins in the same cluster. Finally, the characterization of sequences from clusters that are large is preferred because they are more likely to add information about a larger number of organisms and therefore give broader biological insight. Thus, a third measure of annotation thoroughness is the number of proteins in the clusters that have at least one annotated sequence (column 3 in Table 1).

The analysis of the annotation pooled by cluster can be used in combination with the taxonomic distribution of their members to identify large clusters without an experimentally characterized member but with interesting taxonomic patterns. As a proof of concept, a web query tool to retrieve clusters by taxonomic and annotation properties has been implemented (available as supplementary information online). Using this system, it is possible to deduce that 219 clusters include members from *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Xenopus laevis*, *Drosophila melanogaster* and *Caenorhabditis elegans*, of which five have neither literature nor protein domain annotations. One such example is depicted in Fig 3.

As stated above, a high fraction of clusters are annotated with some domain information; however, this information describes only a fragment of the protein sequence. What about the coverage in terms of protein length? The analysis presented here—available

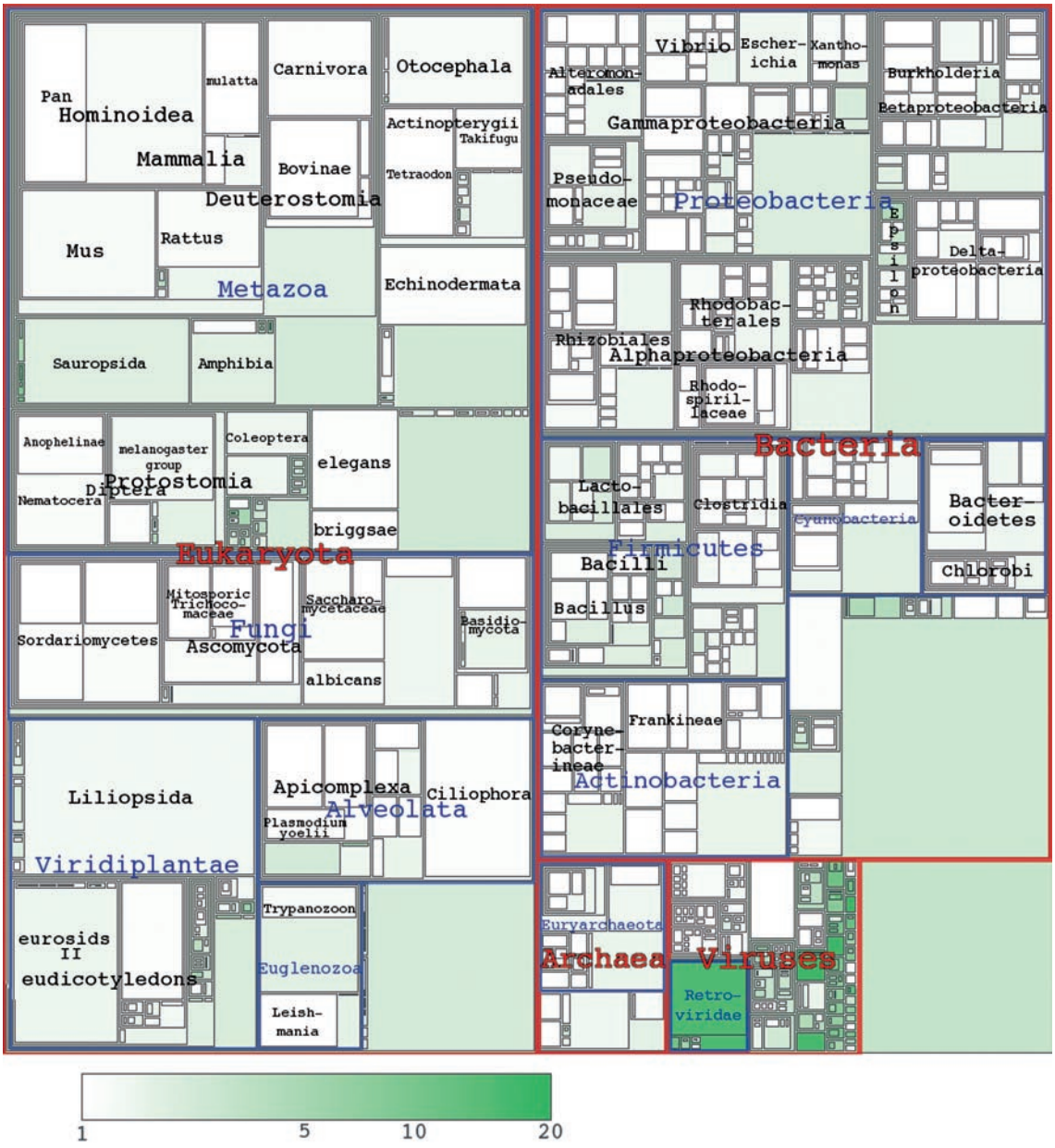

Fig 2 | Taxonomic distribution of all protein clusters from UniRef100. Treemap visualization (Shneiderman, 1992) of the taxonomic distribution of the 1.35 million clusters obtained by clustering UniRef100 release 8.5 (September 2006). The size of the boxes is proportional to the number of clusters at that taxonomic node; the colour intensity indicates the average cluster size (from 1, white, to 20, dark green, in a logarithmic scale). The treemap was generated from the full list of all clusters. For each cluster, the most general taxonomic node in common was identified. The aggregate number of nodes was then calculated for each position in the taxonomic tree. The 1,000 taxonomic nodes with the highest cumulative count—all clusters at that node and below—were selected for representation on the treemap. To simplify the diagram, only those taxonomic nodes that were 90% smaller than their closest represented ancestor node were shown. The resulting set of taxonomic nodes was rendered using a modified version of Treemap-0.2. To emphasize interesting features of the diagram, labels were added manually. A similar graph is available online from http://www.ogic.ca/projects/clusters/ in which taxa labels can be observed by mouse hovering, and boxes are linked to the corresponding taxonomic database entry at the National Center for Biotechnology Information. All underlying data are provided in Table S1 available at: http://www.ogic.ca/projects/clusters/sorted_allcluster_taxonomy_8.5.zip.

in full in the supplementary Methods section online—indicates that although 53% of the cluster representatives, with an average length of 392 amino acids (aa), have some domain information, only 32% of their aggregate sequence length is covered by these domains at present. That is, 168 million of 531 million aa matched to Pfam domains, with an average length of 160 aa and a total number of 1 million matches to 8,200 different Pfam domains. One reason for this is that not all domains have yet been defined. Also, large parts

Table 1 | Protein annotation by feature in UniRef100 v8.5

	Sequences	Clusters	Sequences in annotated clusters	Seq./Clust.	Gain
Three dimensional structure	10,707 (0.30%)	8,741 (0.65%)	217,399 (6.1%)	24.9	20.3
Bibliography	207,446 (5.8%)	106,334 (7.8%)	1,059,819 (30%)	10.0	5.1
Function	1,727,707 (48%)	498,124 (37%)	2,303,804 (65%)	4.6	1.33
Domain	2,523,240 (71%)	723,518 (53%)	2,661,121 (75%)	3.7	1.05
Total	3,562,626	1,354,861	3,562,626	2.6	1.00

Sequences: Protein sequences annotated. Clusters: Clusters containing at least one annotated protein (annotated clusters). Sequences in annotated clusters: Number of sequences in annotated clusters. Seq./Clust.: Average number of sequences in annotated clusters (column 3 divided by column 2). Gain: Absolute coverage gain by clustering (column 3 divided by column 1). All annotations are provided in Table S2 available at: http://www.ogic.ca/projects/clusters/uniref100_8_5_cluster_annotations.zip.

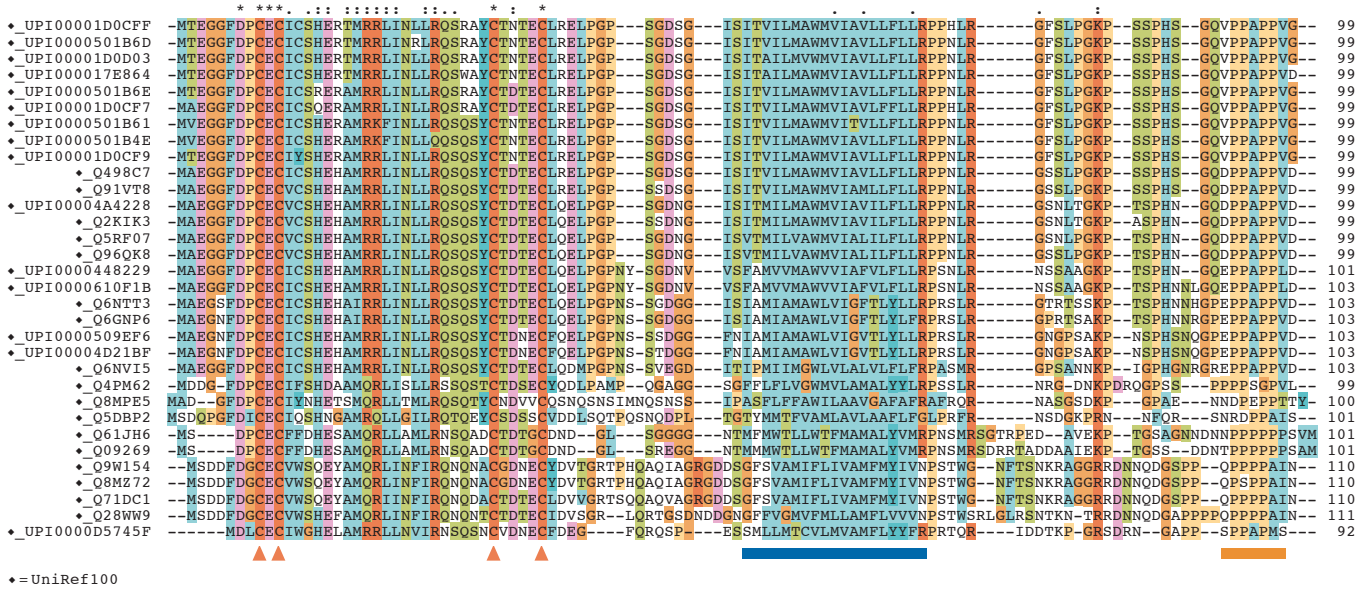

Fig 3 | Sequence alignment of members of cluster UniRef100_Q28WW9. The cluster UniRef100_Q28WW9 contains 32 proteins including the products of human *C4orf34*, mouse *1110003E01Rik* and fruit fly *AT28250p* hypothetical genes, as well as proteins from other metazoa. A PSI-BLAST search of the NCBI's protein database using the UniRef100_Q28WW9 sequence (cluster leader, from *Drosophila pseudoobscura*) converged to a similar set of sequences. The 32 members of cluster UniRef100_Q28WW9 were aligned using ClustalW (Thompson *et al*, 1994). Matches to a transmembrane region (predicted using Phobius; Kall *et al*, 2004) and a carboxy-terminal proline-rich region (obtained using BiasViz; Huska *et al*, 2007) are indicated at the bottom with a blue and an orange bar, respectively. Conserved cysteines—indicated with red triangles—can be observed in the N-terminal region, whereas there are none in the C-terminal region. This suggests that the N terminus of the protein is extracellular and the C terminus cytoplasmic, a conclusion reached also by the Phobius server. Current versions of the databases (June 2007) did not include specific functional or bibliographic information for any member of this cluster. Thus, this family, which represents a small transmembrane protein conserved in metazoans, constitutes a potentially interesting target of experimental verification. NCBI, National Center for Biotechnology Information.

of protein sequences are not amenable to definition in terms of evolutionarily conserved domains because they bear no evolutionary pressure for sequence conservation, which is generally due to the absence of strong structural constraints on their sequences, such as coiled coils, transmembrane helices and low-complexity regions. Our estimation is that only 53%—281 million of 531 million aa—of the aggregated sequence of the cluster representatives might be amenable to definition by homology to conserved domains. This figure agrees well with previous predictions (Wootton, 1994). Thus, the total sequence length that is now covered by homology to domains is 60% of the total sequence that could possibly be covered.

Following our previous conjecture, assuming the estimated value of 5 million for the total number of distinct proteins in the Earth's proteome, an approximate total length of the Earth's proteins covered by structural domains would be 1,039 million aa (that is, 5 million proteins x 392 aa average length x 53% of sequence covered by conserved domains). This would correspond to approximately 6.5 million instances of domains (that is, 1,039 million aa/160 aa per domain), which is 6.5 times the current amount. As of February 2007, there are approximately 8,200 domains in the Pfam database. Assuming that the current distribution of domain occupancy (supplementary Fig S1 online) scales up with further domain characterization, the total

number of domains in the Earth's proteome could be estimated to be 20,900 (supplementary Methods see online). By this estimate, the current number of domains in the Pfam domain database would be 39% of the total protein domains on Earth.

Conclusion

We have expanded the idea that the discovery of protein families must slow down as sequence space coverage increases (Vitkup *et al*, 2001). Similar to previous studies (Marsden *et al*, 2006; Raes *et al*, 2007), we have analysed this trend assuming that clusters of similar sequences are good representatives of protein families. We note that any study in which proteins grouped by similarity are assumed to be somehow functionally equivalent is subject to some level of oversimplification. Some members of fast diverging protein families might be undetected by the clustering method, and the amount of mutations that can make two proteins different cannot be measured by percentage of identity, because that will depend on the functions under consideration and the particular mutations that occurred. However, in large-scale analyses such as the ones cited, simplifying assumptions are a necessary evil. In the study presented here, a clustering procedure tailored specifically for efficiency allowed us to study all historical versions of the complete protein database and, thus, to observe a trend predicted in 2001 (Vitkup *et al*, 2001) for the first time. We have adopted a conservative strategy in which the sequences clustered by similarity must have similar length and therefore the same domain organization. Although this approach implies that some very large clusters might be split, because these large clusters are extremely rare, this affects our conclusions only minimally. Another caveat is that we are assuming that current and historical sequencing rates can be extrapolated to the next few years, and this might not be the case. For example, new sequencing technologies might unexpectedly accelerate the production of new sequences, or new socio-economic factors could direct sequencing to particular environments and organisms. In summary, the large-scale analysis presented here is necessarily affected by a degree of computational uncertainty. However, we believe that these results will be correct to within an order of magnitude. In any case, we have presented a protocol that will allow refinement of these predictions as completion of the Earth's proteome approaches.

The novel view of the protein data set through 'full-length similar' clusters provides a method to monitor progress in sequencing and annotation. This allows not only the estimation of the number of proteins on Earth and the time that will be required to discover them, but also a plan for future sequencing, experimental characterization, and annotation efforts. The analysis of protein annotations suggests that the methods used for annotation are conservative and infer annotations from a very small amount of experimental information. New strategies are necessary to derive experimental information from proteins selected according to the structure of the clustered database in order to optimally cover the largest uncharacterized families.

A close collaboration between computational biologists, engineers—to develop novel sequencing technologies—and microbiologists—to hunt for the species that will close the gaps in the Earth's proteome—will be required to take advantage of the information provided by this method. Such an effort will significantly hasten the sequencing and characterization of the Earth's proteome.

Supplementary information is available at *EMBO reports* online (http://www.emboreports.org)

ACKNOWLEDGEMENTS

All the members of our Bioinformatics group at the Ottawa Health Research Institute are thanked for fruitful discussions. This work has been supported with funds from the Canadian Foundation for Innovation, the Ontario Innovation Trust and the Ontario Research and Development Challenge Funds. M.A.A.-N. is a recipient of a Canada Research Chair in Bioinformatics.

REFERENCES

Adam GC, Sorensen EJ, Cravatt BF (2002) Chemical strategies for functional proteomics. *Mol Cell Proteomics* **1:** 781–790

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634

Bairoch A, Boeckmann B (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **19 (Suppl):** 2247–2249

Bairoch A *et al* (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res* **35:** D193–D197

Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C (1995) Challenging times for bioinformatics. *Nature* **376:** 647–648

Finn RD *et al* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* **34:** D247–D251

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99–113

Fleischmann RD *et al* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512

Fraser CM *et al* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403

Gaasterland T (1998) Structural genomics taking shape. *Trends Genet* **14:** 135

Huska MR, Buschmann H, Andrade-Navarro MA (2007) BiasViz: visualization of amino acid biased regions in protein alignments. *Bioinformatics* [doi:10.1093/bioinformatics/btm489]

Huynen MA, Bork P (1998) Measuring genome evolution. *Proc Natl Acad Sci USA* **95:** 5849–5856

Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338:** 1027–1036

Koonin EV, Mushegian AR (1996) Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr Opin Genet Dev* **6:** 757–762

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34:** D257–D260

Marsden RL, Lee D, Maibaum M, Yeats C, Orengo CA (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res* **34:** 1066–1080

Ottesen EA, Hong JW, Quake SR, Leadbetter JR (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* **314:** 1464–1467

Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* **276:** 734–740

Perez-Iratxeta C, Andrade-Navarro MA, Wren JD (2007) Evolving research trends in bioinformatics. *Brief Bioinform* **8:** 88–95

Raes J, Harrington ED, Singh AH, Bork P (2007) Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol* **17:** 362–369

Rondon MR *et al* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66:** 2541–2547

Sanger F (2001) The early days of DNA sequences. *Nat Med* **7:** 267–268

Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B (2001) Industrial biocatalysis today and tomorrow. *Nature* **409:** 258–268

Shneiderman B (1992) Tree visualization with tree-maps: a 2-dimensional space filling approach. *ACM Trans Graph* **11:** 92–99

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23:** 1282–1288

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680
Torsvik V, Ovreas L, Thingstad TF (2002) Proka;ryotic diversity—magnitude, dynamics, and controlling factors. *Science* **296:** 1064–1066
Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37–43
Venter JC *et al* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74
Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. *Nat Struct Biol* **8:** 559–566
Wheeler DL *et al* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35:** D5–D12
Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95:** 6578–6583
Wootton JC (1994) Sequences with 'unusual'amino acid compositions. *Curr Opin Struct Biol* **4:** 413–421
Yooseph S *et al* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5:** e16

Carolina Perez-Iratxeta

Gareth Palidwor

Miguel A. Andrade-Navarro